# Interpretable Encoding of Densities using Possibilistic Logic

**Ondřej Kuželka**[1] and **Jesse Davis**[2] and **Steven Schockaert**[3]

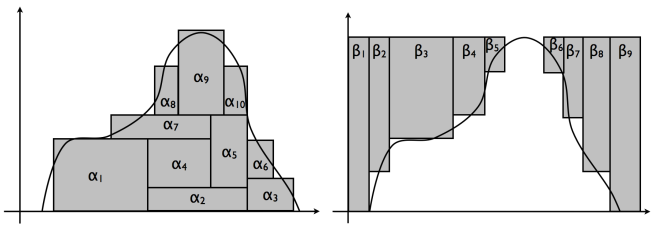
**Abstract.** Probability density estimation from data is a widely studied problem. Often, the primary goal is to faithfully mimic the underlying empirical density. Having an interpretable model that allows insight into why certain predictions were made is often of secondary importance. Using logic-based formalisms, such as Markov logic, can help with interpretability, but even in Markov logic it can be difficult to gain insight into a model's behavior due to interactions between the logical formulas used to specific the model. This paper explores an alternative approach to representing densities that makes use of possibilistic logic. Concretely, we propose a novel way to transform a learned density tree into a possibilistic logic theory. An advantage of our transformation is that it permits performing both MAP and, surprisingly, marginal inference, with the converted possibilistic logic theory. At the same time, we still retain the benefits conferred by using possibilistic logic, such as the ability to compact the theory and the interpretability of the model.

## 1 INTRODUCTION

A key machine learning task is learning to compactly represent a probability density from a given set of examples. The resulting distributions can be useful for answering a large variety of queries. While many sophisticated approaches exist for this task [7, 21, 27, 23, 25, 26] usually the focus is on ensuring that the model as accurately as possible captures the empirical distribution. Often, this comes at the expense of interpretability, which makes it difficult to gain insight into a model's predictions or to refine it based on feedback from users or experts.

A natural approach for representing densities in an interpretable way is to describe them using logical formulas (or in a framework which is close to logic such as Bayesian networks). Figure 1(a) shows the idea underlying Markov logic [26], one of the most popular logic-based frameworks for modelling densities. This example approximates the density using the propositional formulas $\alpha_1, ..., \alpha_{10}$, each of which has a corresponding weight $w_i$. In Figure 1(a), the height of each box is related to the weight associated with the corresponding formula, and the width represents its set of models. The probability of a given possible world is defined to be proportional to the exponentiated sum of the weights of the satisfied formulas. Still, it can be difficult to grasp the intuitive meaning of the weights associated with the formulas. In the considered example, for instance, $\alpha_1$ has one the highest weights, which could incorrectly give the impression that models of $\alpha_1$ are highly probable.

This paper explores an alternative approach, also based on weighted logical formulas, which is illustrated in Figure 1(b). Here, a weight's meaning is clear as there is a more explicit relationship between it and the probability of the corresponding possible worlds: a formula $\beta$ with weight $1 - p$ means that its models can have at most a probability of $p$. Thus, weighted formulas are seen as constraints that act on the set of possible worlds. In Figure 1(b), the most probable worlds are exactly those that satisfy $\neg\beta_1, ..., \neg\beta_9$. Similarly, if the weight associated with formulas $\beta_5$ and $\beta_6$ is $1 - p_1$, then the worlds whose probability is higher than $p_1$ are exactly those who satisfy $\neg\beta_1, ..., \neg\beta_4, \neg\beta_7, ..., \neg\beta_9$.




(a) Additive approximation (b) Constraint based approximation

**Figure 1.** Two ways of approximating a density using logical formulas.

To the best of our knowledge, we present the first such constraint-based representation of probability densities in a logical setting. Specifically, the approach we propose consists of estimating a density tree [25] from a given set of examples, and then converting it into a possibilistic logic theory [8]. Possibilistic logic is a well-known representation and reasoning formalism that supports non-monotonic inferences, which is a requirement if we are to model probability densities. At the same time, it remains close to classical logic, which helps with interpretability and means that off-the-shelf SAT solvers enable highly efficient reasoning. We present several novel ways to perform this transformation. We show that it is possible to perform both MAP and, surprisingly, marginal inference, with the converted possibilistic logic theory. Using possibilistic logic confers several advantages. First, because possibilistic logic remains close to classical logic, we can exploit logical inference to reduce the size of the learned theories, sometimes leading to theories that are exponentially smaller than the corresponding density trees. Second, we can improve the quality of learned possibilistic logic theories by taking into account feedback or input from experts. To this end, we provide an

[1] School of Computer Science and Informatics, Cardiff University, UK, email: KuzelkaO@cardiff.ac.uk
[2] Department of Computer Science, KU Leuven, Belgium, email: jesse.davis@cs.kuleuven.be
[3] School of Computer Science and Informatics, Cardiff University, UK, email: SchockaertS1@cardiff.ac.uk

algorithm for retraining the weights of the possibilistic logic theory that preserves the expert's modifications. Finally, the resulting theories should be more interpretable.

An implementation of the methods described in this paper is available online[4].

## 2 BACKGROUND

### 2.1 Possibilistic logic

**Syntax** A theory in possibilistic logic is a set of formulas $\{(\alpha_1, \lambda_1), ..., (\alpha_n, \lambda_n)\}$ with each $\alpha_i$ a propositional formula and $\lambda_i$ a certainty weight in $]0, 1]$. The standard inference relation in possibilistic logic follows the principle of the weakest link, i.e. $\{(\alpha_1, \lambda_1), ..., (\alpha_n, \lambda_n)\} \vdash (\alpha^*, \lambda^*)$ if the following entailment holds in classical logic: $\{\alpha_i \,|\, \lambda_i \geq \lambda^*\} \models \alpha^*$. In other words, we can derive $(\alpha^*, \lambda^*)$ iff we can classically derive $\alpha^*$ without using formulas whose certainty weight is strictly less than $\lambda^*$.

The $\lambda$-cut of a possibilistic logic theory $\Theta$ is defined as the classical theory $\Theta^\lambda = \{\gamma | (\gamma, \mu) \in \Theta \text{ and } \mu \geq \lambda\}$. A non-monotonic consequence relation $\vdash_{poss}$ can be defined for possibilistic logic as follows. Consider a possibilistic logic theory $\Theta$ and formula $\alpha$. Let $\lambda^*$ be the highest certainty value for which $\{\alpha\} \cup \Theta^{\lambda^*}$ is inconsistent (and $\lambda^* = 0$ if there is no such certainty value). Then $(\Theta, \alpha) \vdash_{poss} \beta$ iff the entailment $\{\alpha\} \cup \{\alpha_i \,|\, (\alpha_i, \lambda_i) \in \Theta, \lambda_i > \lambda^*\} \models \beta$ holds in classical logic. Note that all formulas whose certainty weight is at most $\lambda^*$ are ignored, even if they are unrelated to any inconsistency in $\Theta$. This is known as the drowning effect.

**Semantics** The semantics of possibilistic logic are defined in terms of possibility distributions. A possibility distribution, in this context, is a mapping $\pi$ from the set of possible worlds $\Omega$ to $[0, 1]$. A possibility distribution induces two uncertainty measures: the possibility measure $\Pi$ and its dual $N$, defined for $A \subseteq \Omega$ as [30, 10]:

$$\Pi(A) = \max_{a \in A} \pi(a) \qquad N(A) = 1 - \Pi(\Omega \setminus A)$$

We will also write $N(\alpha)$ as an abbreviation for $N([\![\alpha]\!])$, where $\alpha$ is a propositional formula and $[\![\alpha]\!]$ is its set of models. Intuitively, $\Pi(\alpha)$ reflects the degree to which $\alpha$ is compatible with our available beliefs, while $N(\alpha)$ reflects the degree to which $\alpha$ is considered certain. At the semantic level, the possibilistic logic theory $\Theta = \{(\alpha_1, \lambda_1), ..., (\alpha_n, \lambda_n)\}$ corresponds to the possibility distribution $\pi$ defined by ($\omega \in \Omega$):

$$\pi(\omega) = 1 - \max\{\lambda_i \,|\, \omega \not\models \alpha_i\} \tag{1}$$

where we assume $\max \emptyset = 0$. It is easy to see that, for $N$ the necessity measure induced by $\pi$, it holds that $N(\alpha_i) \geq \lambda_i$. Moreover, it can be shown that $\Theta \vdash (\alpha, \lambda)$ iff $N(\alpha) \geq \lambda$ [20].

### 2.2 Density estimation trees

A density estimation tree[5] is a rooted directed binary tree in which internal nodes are labelled by propositional variables (attributes) and leaves are labelled by real numbers ("densities"). $\mathsf{Nodes}(T)$ denotes the set of the nodes in tree $T$ and $\mathsf{Leaves}(T)$ the set of its leaves. Edges are labelled by 0 (false) or 1 (true). A path from the root to a leaf is called a branch. We will represent branches by conjunctions. For instance, let $a_1, 0, a_{i_2}, 1, \ldots, 1, a_{i_{k-1}}, 0$ be the sequence of labels of internal nodes and edges corresponding to a branch from the root $N_1$, labelled with propositional variable $a_1$, to a leaf $N_{i_k}$. Then the conjunction corresponding to this branch is $\neg a_1 \wedge a_{i_2} \cdots \wedge \neg a_{i_{k-1}}$. We call branches paths or conjunctions interchangeably. A density estimation tree defines a probability distribution on possible worlds where the probability of a world $\omega$ is given by the label of the leaf of the unique branch $B$ which is consistent with $\omega$ (i.e. such that $\omega \models B$). Importantly, density estimation trees can be learned efficiently from data [25].

## 3 REPRESENTING DENSITY TREES IN POSSIBILISTIC LOGIC

Next, we show how we can transform a density tree into a possibilistic logic theory. Surprisingly, this transformation permits computing marginal probabilities from the possibilistic logic theory.

### 3.1 Transforming density trees

We start with a basic transformation which is similar in spirit to transforming a decision tree into a CNF formula.

**Transformation 1.** *Let $T$ be a density estimation tree. Let $\mathcal{B} = \{B_1, \ldots, B_k\}$ be the set of all branches of the tree, represented as conjunctions, and let $p_1, \ldots, p_k$ be the estimated probabilities of worlds consistent with the respective branches, i.e. if $\omega \models B_i$ then $p(\omega) = p_i$. We define the possibilistic logic theory corresponding to $T$ as $\Theta_T = \{(\neg B_i, 1 - p_i) | B_i \in \mathcal{B}\}$.*

**Proposition 1.** *If $T$ is a density estimation tree and $\Theta_T$ is its possibilistic logic theory constructed by Transformation 1, then for all possible worlds $\omega$, it holds that $p(\omega) = \pi(\omega)$, where $p(.)$ is the probability given by $T$ and $\pi(.)$ is the possibility distribution associated with $\Theta_T$.*

*Proof.* Let $\omega$ be a possible world, $T$ be a density estimation tree and $\mathcal{B} = \{B_1, \ldots, B_k\}$ be the set of all its branches, represented as conjunctions. Let $B^* \in \mathcal{B}$ be a branch consistent with $\omega$ and $p^*$ be the respective density, i.e. $\omega \models B^*$. Clearly, there can be only one such branch as all the branches in $T$ are mutually exclusive. Likewise, the only rule $(\alpha', \lambda') \in \Theta$ which is not satisfied in $\omega$ is $(\neg B^*, 1 - p^*)$. By (1), we therefore have $\pi(\omega) = 1 - (1 - p^*) = p^*$. It follows that for any $\omega \in \Omega$ we have $\pi(\omega) = p(\omega)$. □

**Remark 1.** *Transformation 1 works in time $O(|\mathsf{Nodes}(T)|^2)$. A possibilistic logic theory constructed by Transformation 1 from a density estimation tree $T$ has at most $|\mathsf{Leaves}(T)|$ rules and its size, i.e. the sum of the lengths of its rules, is bounded by $|\mathsf{Nodes}(T)|^2$. Figure 2 shows an example of a tree, whose possibilistic logic representation is quadratic. Specifically, the size of the possibilistic logic theory constructed from such a tree of depth $k$ (which has* size $S = 2k - 1$ *nodes) by Transformation 1 is of order $O(k^2)$ which is also of order $O(S^2)$.*

In Section 4.1 we will show how the size of the constructed possibilistic logic theories can be reduced, often significantly. As we will see, there are cases where the reduced possibilistic logic theories are exponentially smaller than the trees from which they were created. Next we give an example of applying Transformation 1.

[4] https://github.com/supertweety/

[5] In this paper we only consider density estimation trees involving Boolean variables. In general, density estimation trees can also define probability densities in continuous domains.

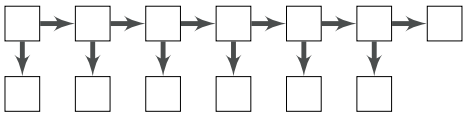


**Figure 2.** A tree whose possibilistic logic representation is quadratic.

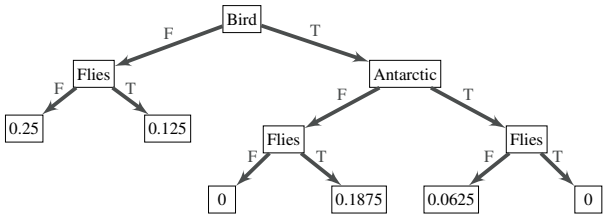



**Figure 3.** A density estimation tree from Example 1.

**Example 1.** *Applying Transformation 1 to the density estimation tree $T$ in Figure 3 yields the following possibilistic logic theory*

$$\begin{aligned}\Theta = \{&(\neg \textit{bird} \vee \textit{antarctic} \vee \textit{flies}, 1),\\ &(\neg \textit{bird} \vee \neg \textit{antarctic} \vee \neg \textit{flies}, 1),\\ &(\neg \textit{bird} \vee \neg \textit{antarctic} \vee \textit{flies}, 0.9375),\\ &(\textit{bird} \vee \neg \textit{flies}, 0.875),\\ &(\neg \textit{bird} \vee \textit{antarctic} \vee \neg \textit{flies}, 0.8125),\\ &(\textit{bird} \vee \textit{flies}, 0.75)\}.\end{aligned}$$

*Notice that we have, e.g.* $(\Theta, \emptyset) \vdash_{poss} \neg \textit{bird}$, $(\Theta, \{\textit{bird}\}) \vdash_{poss} \textit{flies} \wedge \neg \textit{antarctic}$.

**Remark 2.** *In the possibilistic logic theory obtained by Transformation 1, the formula with the lowest weight always drowns, i.e. it is inconsistent with the other formulas and will thus never play a role in the evaluation of $\vdash_{poss}$. In many cases, it is therefore possible to remove this formula from the theory. However, if we want to use the possibilistic logic theory for probabilistic inference, as in Section 3.2, then we must keep the lowest level in the theory (although we can replace it by $\bot$) in order to keep the information about the numerical value of the probability of the most probable worlds.*

Transformation 1 differs from the standard probability-possibility transformation [11]. For completeness, we present the syntactical counterpart of the standard probability-possibility transformation in Appendix A.1. In fact, both transformations induce the same ranking of possible worlds (since both are identical to the ranking induced by the probability distribution). Therefore, any classical formula $\alpha$ which can be derived using the possibilistic entailment operator $\vdash_{poss}$ from a theory obtained by one of the transformations can also be derived from the theory obtained by the other transformation.

It is beneficial for scalability to simplify the possibilistic theory as much as possible already while performing the transformation, or at least without having to check logical entailment. To this end, we describe a simple modification of Transformation 1 which results in smaller possibilistic logic theories.

**Transformation 2.** *Let $T$ be a density estimation tree. Let $\mathcal{B} = \{B_1, \ldots, B_k\}$ be the set of all branches of the tree, represented as conjunctions, and let $p_1, \ldots, p_k$ be the estimated probabilities of worlds consistent with the respective branches, i.e. if $\omega \models B_i$ then $p(\omega) = p_i$. Let us assume w.l.o.g. that $p_i \geq p_{i+1}$. The resulting possibilistic logic theory then consists of possibilistic logic formulas $(\neg B'_i, 1 - p_i)$ where each $B'_i$ is a conjunction and is obtained as follows:*

- *Without loss of generality, let $B_i = b^i_1 \wedge \cdots \wedge b^i_{j_i}$ where its conjuncts are ordered so that the node labeled by $b^i_s$ is closer to the root of $T$ than the node labeled by $b^t$ whenever $s < t$.*
- *Let $B'_i = b^i_1 \wedge \cdots \wedge b^i_r$ be the minimal prefix of $B_i$ such that there is no $j < i$ such that $B_j$ contains $B'_i$ as a prefix.*

**Proposition 2.** *Transformation 1 and Transformation 2 produce equivalent possibilistic logic theories.*

*Proof.* Let us denote by $\Theta_A$ the result of Transformation 1 and by $\Theta_B$ the result of Transformation 2. To prove this proposition it is enough to show that $\Theta_A$ and $\Theta_B$ correspond to the same possibility distribution. Let $B^A_1, \ldots, B^A_k$ be defined as in Transformation 1 and let $B^B_1, \ldots, B^B_k$ be defined as in Transformation 2. Let $\omega$ be a possible world and let $\neg B^A_{i^*}$ be the only formula from $\Theta_A$ not satisfied in $\omega$ (it follows from the proof of Proposition 1 that there is only one such formula in $\Theta_A$). Clearly the respective formula $\neg B^B_{i^*}$ cannot be satisfied in $\omega$ (because $\neg B^B_{i^*}$ implies $\neg B^A_{i^*}$). Let $1 - p_{i^*}$ be the weight of $\neg B^A_{i^*}$ in $\Theta_A$ (equal to the weight of $\neg B^B_{i^*}$ in $\Theta_B$). What we need to show is that $\neg B^B_{i^*}$ has the highest weight among the formulas from $\Theta_B$ which are not satisfied in $\omega$. It follows from the way Transformation 2 works that any such formula would necessarily have to be a prefix of $\neg B^B_{i^*}$ (i.e. a clause consisting of the first $r$ literals of $\neg B^B_{i^*}$). This is because every clause in $\Theta_B$ is a prefix of some clause in $\Theta_A$ and at most one clause from $\Theta_A$ can be falsified in any possible world $\omega$ at the same time. But then it follows that $\neg B^B_{i^*}$ must have the highest weight among the falsified formulas because, by construction, there is no clause $\neg B^B_j$ with $j > i^*$ which is a prefix of $\neg B^B_{i^*}$. It follows that $\pi_A(\omega) = \pi_B(\omega)$, where $\pi_A$ and $\pi_B$ are the possibility distributions corresponding to $\Theta_A$ and $\Theta_B$, respectively. □

The next example illustrates the use of Transformation 2.

**Example 2.** *Let us consider the same density estimation tree as in Example 1 (shown in Figure 3). The branches of the tree, represented as conjunctions, are: $B_1 = \neg \textit{bird} \wedge \neg \textit{flies}$, $B_2 = \textit{bird} \wedge \neg \textit{antarctic} \wedge \textit{flies}$, $B_3 = \neg \textit{bird} \wedge \textit{flies}$, $B_4 = \textit{bird} \wedge \textit{antarctic} \wedge \neg \textit{flies}$, $B_5 = \textit{bird} \wedge \textit{antarctic} \wedge \textit{flies}$, and $B_6 = \textit{bird} \wedge \neg \textit{antarctic} \wedge \neg \textit{flies}$. The corresponding conjunctions $B'_i$ are then (we omit $B'_1$ considering Remark 2):*

$$\begin{aligned} B'_2 &= \textit{bird}, & B'_3 &= \neg \textit{bird} \wedge \textit{flies}\\ B'_4 &= \textit{bird} \wedge \textit{antarctic}, & B'_5 &= \textit{bird} \wedge \textit{antarctic} \wedge \textit{flies}\\ B'_6 &= \textit{bird} \wedge \neg \textit{antarctic} \wedge \neg \textit{flies} \end{aligned}$$

*Applying Transformation 2 yields the following possibilistic logic theory: $\Theta' = \{(\neg \textit{bird} \vee \textit{antarctic} \vee \textit{flies}, 1), (\neg \textit{bird} \vee \neg \textit{antarctic} \vee \neg \textit{flies}, 1), (\neg \textit{bird} \vee \neg \textit{antarctic}, 0.9375), (\textit{bird} \vee \neg \textit{flies}, 0.875), (\neg \textit{bird}, 0.8125)\}$.*

## 3.2 Answering queries

In this section, we discuss how different types of queries about a given density can be answered by using the possibilistic logic theory obtained from Transformation 1 or 2. The most natural kinds of

queries to consider, in the context of possibilistic logic, are maximum a posteriori (MAP) queries, as these only depend on the ordering of the possible worlds. In particular, we consider the following MAP inference relation [13]:

$$(T, \alpha) \vdash_{MAP} \beta \quad \text{iff} \quad \forall \omega \in \max(T, \alpha) : \omega \models \beta$$

where $T$ is a density tree, $\alpha$ and $\beta$ are propositional formulas and $\max(T, \alpha)$ is the set of most probable models of $\alpha$, w.r.t. the probability distribution induced by $T$. The next proposition shows that, for possibilistic logic theories obtained using the introduced transformations, whatever can be derived using the $\vdash_{MAP}$ relation from the tree can also be derived using $\vdash_{poss}$ from the respective possibilistic logic theory, and vice versa.

**Proposition 3.** *If $T$ is a density estimation tree and $\Theta_T$ is the possibilistic logic theory constructed using Transformation 1 or 2, or using Transformation 4 described in the appendix, then for any formulas $\alpha$ and $\beta$ it holds that*

$$(T, \alpha) \vdash_{MAP} \beta \quad \textit{iff} \quad (\Theta_T, \alpha) \vdash_{poss} \beta.$$

*Proof.* It is sufficient to show that the ranking of possible worlds induced by the probability $p(.)$ defined by the density estimation tree is the same as the rankings of possible worlds induced by the possibility distributions $\pi(.)$ corresponding with the theories that are obtained by the three transformations. For Transformation 1 this follows from Proposition 1. For Transformation 2, this follows from Proposition 1 and Proposition 2. For Transformation 4, this follows from the so-called order preservation principle, which is known to hold for the probability-possibility transformation (see [12]). □

One important advantage of Transformation 1 and 2 over the standard probability-possibility transformation is that it permits computing marginal probabilities directly from the possibilistic logic theory. In particular, if $\Theta_T$ is a possibilistic logic theory obtained by Transformation 1 or 2 then the probability of a formula $\alpha$ is

$$P(\alpha) = \sum_{\omega : \omega \models \alpha} \pi(\omega). \tag{2}$$

The convenience of the possibilistic logic encoding lies in the fact that the sum in (2) can easily be computed using model counting as follows.

- Let $\Theta$ be a possibilistic logic theory obtained by Transformation 1 or 2 and let $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_k\}$ be the set of weights appearing in $\Theta$, sorted in increasing order. Let us set $\lambda_{k+1} = \infty$ (for convenience of notation below). For every $\lambda \in \Lambda$ let us define $M^\alpha_\lambda = \{\omega | \omega \models \Theta^\lambda \cup \{\alpha\}\}$, with $\Theta^\lambda$ the $\lambda$-cut of $\Theta$ as before. In other words, $|M^\alpha_\lambda|$ is the "model count" of $\Theta^\lambda \cup \{\alpha\}$.
- We find:

$$P_\Theta(\alpha) = \sum_{i=1}^{k} (1 - \lambda_i) \cdot \left( |M^\alpha_{\lambda_{i+1}}| - |M^\alpha_{\lambda_i}| \right). \tag{3}$$

**Proposition 4.** *If $\Theta$ is a possibilistic logic theory whose corresponding possibility distribution $\pi$ can be interpreted as a probability distribution (i.e. $\sum_\omega \pi(\omega) = 1$). Let $P$ be the probability measure induced by $\pi$. It holds that $P(\alpha) = P_\Theta(\alpha)$.*

*Proof.* We need to show that $P_\Theta(\alpha) = \sum_{\omega : \omega \models \alpha} \pi(\omega)$. Let $\Lambda$ and $\lambda_{k+1}$ be defined as above. We have

$$\sum_{\omega : \omega \models \alpha} \pi(\omega) = \sum_{\omega : \omega \models \alpha} 1 - \max\{\lambda | (\gamma, \lambda) \in \Theta \text{ and } \omega \not\models \gamma\}$$

$$= \sum_{i=1}^{k} (1 - \lambda_i) \cdot |\{\omega \in \Omega | \omega \models \Theta^{\lambda_{i+1}} \cup \{\alpha\}\} \setminus$$

$$\{\omega \in \Omega | \omega \models \Theta^{\lambda_i} \cup \{\alpha\}\}| = \sum_{i=1}^{k} (1 - \lambda_i) \cdot \left( |M^\alpha_{\lambda_{i+1}}| - |M^\alpha_{\lambda_i}| \right)$$

where the last equality follows from the fact that $M^\alpha_{\lambda_i} \subseteq M^\alpha_{\lambda_{i+1}}$. □

If $\alpha$ is just a conjunction of literals, and the possibilistic logic theory $\Theta$ is the direct transformation of a density estimation tree, then performing inference in the tree will likely be more straightforward. If the theory $\Theta$ has been modified (e.g., refined by an expert) or $\alpha$ is a more complicated formula than a conjunction of literals, then performing inference directly in the possibilistic logic theory may be more efficient.

Using possibilistic logic inference, we can also easily characterize what is true in all worlds whose probability is above a given threshold $\theta$. In particular, it is easy to see that:

$$[\![\{\alpha \,|\, (\alpha, \lambda) \in \Theta, \lambda > 1 - \theta\}]\!] = \{\omega \,|\, p(\omega) \geq \theta\}$$

where $p$ is the probability distribution associated with the density tree $T$ that was used to construct the possibilistic logic theory $\Theta$, using Transformation 1 or 2. It follows that the set of formulas $\{\alpha \,|\, (\alpha, \lambda) \in \Theta, \lambda > 1 - \theta\}$ exactly characterizes what is true for all worlds whose probability is at least $\theta$. Along similar lines, using model counting as in Eq. 3, we can easily characterize what is true for the $x\%$ most probable worlds, or even the $x\%$ most probable worlds in which some formula $\alpha$ is true.

# 4 IMPROVING LEARNED THEORIES

Learned possibilistic logic theories may be improved in multiple ways, some of which we describe in this section.

## 4.1 Pruning

**Exact pruning** An important advantage of encoding density estimation trees in possibilistic logic is that the resulting theories can be simplified based on logical inference. In particular, a weighted formula $(\alpha, \lambda)$ can be removed from a theory $\Theta$ if $\Theta^\lambda \setminus \{\alpha\} \models \alpha$. We can iteratively identify and remove such formulas, until the theory $\Theta$ is free from redundancies. To further simplify the theory, note that each of the proposed transformations results in a weighted set of clauses. Clearly, we can replace the weighted clause $(a_1 \vee ... \vee a_k, \lambda)$ by the sub-clause $(b_1 \vee ... \vee b_l, \lambda)$, with $\{b_1, ..., b_l\} \subset \{a_1, ..., a_k\}$, if $\Theta^\lambda \models b_1 \vee ... \vee b_l$. This often results in substantially smaller theories, while yielding the same MAP predictions and probability estimates as the initial density trees.

**Example 3.** *Consider a density tree that assigns a uniform non-zero probability to worlds satisfying the formula $(a_1 \vee a_2) \wedge (a_3 \vee a_4) \wedge \cdots \wedge (a_{n-1} \vee a_n)$, and zero probability to the remaining worlds, then such a density tree will be exponentially larger than the respective possibilistic encoding after pruning, which only contains the formulas $(a_1 \vee a_2, \lambda), ..., (a_{n-1} \vee a_n, \lambda)$.*

**Approximate pruning** It is possible to further simplify the possibilistic logic theories if we drop the requirement that the associated possibility distribution should be identical to the probability distribution encoded by the density tree. A particularly convenient way of reducing possibilistic logic theories is to iteratively merge levels with consecutive weights, each time simplifying the newly created level using the exact pruning method outlined above. One possibility is to iteratively merge the levels with the highest weights, which is especially useful for MAP inference, as this reduction only affects the relative ordering of the least probable worlds. Moreover, whatever can be derived from the reduced theory by $\vdash_{poss}$ can also be derived from the original theory, although the converse does not hold in general. For marginal inference, it is necessary to recompute the weight of the new level but that is straightforward.

**Pruning default rules** If we only care about the ordering of the possible worlds, a possibilistic logic theory $\Theta$ may be seen as a compact representation of a default theory, where a default "if $\alpha$ then typically $\beta$" is in this theory if and only if $(\Theta, \{\alpha\}) \vdash_{poss} \beta$. In some cases, e.g. if we want to explain the theories to people without training in logic, it may be preferable to explain what is captured by a given possibilistic logic theory by presenting these default rules instead. A set of short default rules which are implicitly encoded by the possibilistic logic theory can be extracted using a method described in [19]. The resulting set of defaults is usually too large, however. We describe a practically efficient and theoretically sound method, which we also use in the experiments, for pruning sets of default rules in Appendix A.2.

## 4.2 Parameter reestimation

Recall that experts can easily modify a possibilistic logic theory. However, manually modifying a theory obtained using either Transformation 1 or 2 would require us to reestimate the theory's weights if we wanted to use it for computing marginal probabilities (performing MAP inferences does not require retraining the weights). However, we do not want this retraining to override an expert's modifications. Therefore, we require that the reestimated weights have the same relative ordering as the original weights. This ensures that everything that can be derived by MAP from the original theory can also be derived from the theory with the reestimated parameters (but the probabilities of the possible worlds will be different). Imposing this restriction makes reestimating the parameters a non-trivial problem, for which we present a solution in this subsection.

Let $E$ be a multiset of examples which we want to use to reestimate the parameters. Let $\Theta$ be a possibilistic logic theory, let $\Lambda = \{\lambda_1, \ldots, \lambda_k\}$ be the set of weights in $\Theta$, ordered increasingly, let $\lambda_{k+1} = \infty$ and let $P_\Theta$ be given by Eq. 3. A maximum likelihood estimate of the parameters is a solution of the following optimization problem:

- Variables: $\lambda'_1, \lambda'_2, \ldots, \lambda'_{|\Lambda|}$.
- Maximize: $\prod_{\omega \in E} P(\omega) = \prod_{\lambda_i \in \Lambda} (1 - \lambda'_i)^{|E_{\lambda_i}|}$ where $E_\lambda = \{\omega \in E | \lambda = \max\{\lambda' | (\alpha, \lambda') \in \Theta \text{ and } \omega \not\models \alpha\}\}$.
- Subject to:

$$\lambda'_1 < \lambda'_2 < \cdots < \lambda'_{|\Lambda|} \tag{4}$$

$$\sum_{i=1}^{k} (1 - \lambda'_i) \cdot \left( |M^\top_{\lambda_{i+1}}| - |M^\top_{\lambda_i}| \right) = 1 \tag{5}$$

where $M^\top_{\lambda_1}, \ldots, M^\top_{\lambda_j}$ are as in Eq. 3 where we set $\alpha := \top$ (i.e. $\alpha$ is a tautology).

This is a nonlinear optimization problem which can be solved using off-the-shelf[6] techniques of *geometric programming* [5]. In particular, the general *geometric programming problem* is:

- Minimize: $g_0(x)$
- Subject to:

$$g_i(x) \leq 1, i = 1, 2, \ldots, m \tag{6}$$

$$x > 0 \tag{7}$$

where $x = (x_1, \ldots, x_m) \in R^m$ and $g_i$ is a *posynomial*, i.e. $g_i(x) = \sum_{j=1}^{T_i} c_{ij} \prod_{k=1}^{N} x_k^{a_{ijk}}$ with $c_{ij} \geq 0$ and $a_{ijk} \in R$.

We can follow the same strategy used for maximum likelihood parameter estimates of a multinomial distribution with order constraints in [6]. In order to formulate the problem as a geometric programming problem, we first substitute $\lambda'_i := 1 - x$. Then we replace $|E_{\lambda_i}|$ by the relative frequency $|E_{\lambda_i}|/|E|$ and change maximization to minimization by replacing the terms in the maximized product by their reciprocals. We also replace the strict inequalities in Eq. 4 by nonstrict. A solution close to the optimal but with the strict inequalities satisfied can then later be obtained by simply adding and subtracting suitably tiny numbers from the weights. We rewrite each of the nonstrict inequalities $x_i \geq x_{i+1}$ as a posynomial inequality $x_i^{-1} \cdot x_{i+1} \leq 1$. Finally, we also need to replace the equality in Eq. 5 by an inequality $\leq 1$, which clearly does not change the solution in this case.

## 4.3 Ensembles of predictors

Oftentimes, using a model ensemble, which aggregates the predictions of multiple different models, improves modelling performance. Given an ensemble of density estimation trees $T_1, T_2, \ldots, T_n$, we can apply Transformation 1 or Transformation 2 to obtain a possibilistic encoding of the ensemble by constructing possibilistic logic theories $\Theta_1, \Theta_2, \ldots, \Theta_n$ and combining them. We first show how to construct a *weighted* combination of two possibilistic logic theories

**Transformation 3** (Weighted combination of two theories)**.** *Let $\Theta_A$ and $\Theta_B$ be two possibilistic logic theories and let $a$ and $b$, $a+b=1$, be positive real numbers. The weighted combination of $\Theta_A$ and $\Theta_B$ with weights $a$ and $b$ (denoted by $a \cdot \Theta_A \oplus b \cdot \Theta_B$) is the possibilistic logic theory $\Theta_{AB}$ constructed as follows:*

- *For every $(\alpha, \lambda) \in \Theta_A$, add $(\alpha, a \cdot \lambda)$ to $\Theta_{AB}$.*
- *For every $(\beta, \mu) \in \Theta_B$, add $(\beta, b \cdot \mu)$ to $\Theta_{AB}$.*
- *For every pair $(\alpha, \lambda) \in \Theta_A$, $(\beta, \mu) \in \Theta_B$ add $(\alpha \vee \beta, a\lambda + b\mu)$ to $\Theta_{AB}$.*

This transformation actually corresponds to a special case of a combination operator for possibilistic logic theories [3], from which we immediately obtain the following result.

**Proposition 5.** *Let $\Theta_A$ and $\Theta_B$ be possibilistic logic theories and $\pi_A$ and $\pi_B$ be the corresponding possibility distributions. Let $p_A$ and $p_B$ be two probability distributions on possible worlds and let $a$, $b$, $a+b=1$ be positive real numbers. If for all $\omega \in \Omega$, $p_A(\omega) = \pi_A(\omega)$ and $p_B(\omega) = \pi_B(\omega)$ then also $a \cdot p_A(\omega) + b \cdot p_B(\omega) = \pi_{AB}(\omega)$ where $\pi_{AB}$ is possibility distribution corresponding with $a \cdot \Theta_A \oplus b \cdot \Theta_B$.*

[6] Geometric programming problems can be solved using, e.g. the CVX toolkit [15, 14].

To construct a uniform combination of possibilistic logic theories $\Theta_1, \Theta_2, \ldots, \Theta_k$, we can iteratively apply the merging operator $((\ldots(\frac{2}{3}\cdot(\frac{1}{2}\cdot\Theta_1 \oplus \frac{1}{2}\cdot\Theta_2)\oplus\frac{1}{3}\cdot\Theta_3)\oplus\ldots)$. The caveat is that the size of the produced theory may grow exponentially with the number of combined theories. However, this can be mitigated if we allow some imprecision and apply the approximate pruning procedure from Section 4.1 while iteratively building the combination.

# 5 EXPERIMENTS

In this section we experimentally evaluate the proposed methods. We first provide some examples of learned possibilistic logic theories, after which we present a quantitative evaluation in Section 5.2.

## 5.1 Illustrative examples

We start by contrasting the interpretability of a learned possibilistic logic theory with a corresponding MRF.[7] Then we illustrate how interpretability in some cases can be further improved by approximating the possibilistic logic theory using default rules.

**Possibilistic logic and MRFs** First, we use a credit-default dataset [29], where we only consider a subset of the variables for readability. Using the method proposed in this paper, we obtain the following possibilistic logic theory[8]:

$$(\neg \textit{single}, \lambda_0), (\neg \textit{gradSchool}, \lambda_1), (\textit{male}, \lambda_2), (\textit{single} \vee \textit{male}, \lambda_3),$$
$$(\textit{male} \vee \neg \textit{gradSchool}, \lambda_4), (\textit{single} \vee \textit{university}, \lambda_5),$$
$$(\neg \textit{highSchool}, \lambda_6), (\textit{married} \vee \neg \textit{highSchool}, \lambda_7),$$
$$(\neg \textit{otherMaritalStatus}, \lambda_8), (\neg \textit{otherSchool}, \lambda_8)$$

As well as a number of integrity constraints such as $(\neg \textit{single} \vee \neg \textit{married}, 1)$. The theory contains several interesting rules, which capture the properties that hold for typical people who default on their credit. These people typically are not single, did not go to graduate school, and are males. If they are females, then typically they are single, etc. After simplifying the theory with integrity constraints, the most probable worlds have: *married*, *university*, *male*. While these pieces of information provide insight, the main advantage of being able to interpret the model is that we can understand exactly how it arrives at its predictions and potentially debug it by adding or removing rules. We now consider a learned MRF for the same dataset:

$$\begin{aligned} P(\omega) = \frac{1}{Z}\exp(&0.3\cdot \textit{male} + 3.8\cdot \textit{gradSchool} + 4.3\cdot \textit{university} \\ &+3.3\cdot \textit{highSchool} - 3.7\cdot \textit{otherSchool} + 5.0\cdot \textit{married} \\ &+5.1\cdot \textit{single} + 1.6\cdot \textit{otherMaritalStatus} \\ &-1.0\cdot(\textit{male}\wedge\textit{otherSchool}) - 0.2\cdot(\textit{male}\wedge\textit{otherMaritalStatus}) \\ &-1.7\cdot(\textit{otherSchool}\wedge\textit{married}) - 1.6\cdot(\textit{otherSchool}\wedge\textit{single}) \\ &-10.3\cdot(\textit{gradSchool}\wedge\textit{university}) \\ &-9.4\cdot(\textit{gradSchool}\wedge\textit{highSchool}) - 9.8\cdot(\textit{university}\wedge\textit{highSchool}) \\ &-2.9\cdot(\textit{university}\wedge\textit{otherSchool}) - 11.2\cdot(\textit{married}\wedge\textit{single}) \\ &-8.1\cdot(\textit{married}\wedge\textit{otherMaritalStatus}) \\ &-8.1\cdot(\textit{single}\wedge\textit{otherMaritalStatus})) \end{aligned}$$

[7] Recall that propositional Markov logic networks correspond to Markov random fields (MRFs).

[8] Since in this section, we are only interested in MAP inference, we show only symbolic weights $\lambda_0 < \lambda_1 < \ldots$ in the possibilistic logic theory. Note that the same cannot be done for MRFs.

Note that the last five lines correspond to the integrity constraints. Due to the additive combination of the formula's weights, its predictions are encoded intricately via the interaction between the various formulas. As this particular MRF is quite small, it may be possible to gain insight into its preditions with some effort, but with larger theories this quickly becomes impossible.

**Possibilistic logic and default rules** To improve interpretability, possibilistic logic theories can be approximated by sets of default rules (see Section 4.1). To illustrate this idea, Table 1 shows a possibilistic logic theory and its approximated set of default rules for a dataset about the presence of plants in different states of the US and provinces of Canada [16]. For illustrative purposes, we only consider California, Montana, New Mexico and Texas. For example, the default rule $\textit{nm} \mid\!\sim \textit{tx}$ intuitively means that plants found in New Mexico are also usually found in Texas. Such rules can be written in a form that is easy to understand, even for people without training in logic.

## 5.2 MAP inference and marginal inference

We compare our possibilistic logic theories with learned MRFs on both MAP and marginal inference tasks. We have considered the NLTCS, MSNBC, and Plants datasets, which have 16, 17 and 69 propositional variables, respectively. These datasets are divided into train, tune, and test sets. We learned the models on the train sets and report results on held-out test sets. We implemented the possibilistic approach in Java, using SAT4J [4] for SAT solving and RelSat [17] for model counting. For MRFs we used approximate MAP inference and Gibbs sampling from the Libra toolkit [22]. For MAP-inference evaluation, we compare the following models:

**PosLog** Obtained by Transformation 2 and logical simplification
**PosLog (50%)** Compacting PosLog to 50% of its size using the method from Section 4.1
**PosLog (10%)** Compacting PosLog to 10% of its size using the method from Section 4.1
**MRF** The L1 learned models from [21]
**Baseline** A model which predicts all variables to be false

To generate queries, we randomly selected $k$ literals (where $1 \leq k \leq$ number of variables $-1$) for each test example to serve as the evidence and then predicted the most probable assignment for the remaining variables. We measured both accuracy, which is the fraction of examples predicted correctly, and the average Hamming distance between the test-set example and the predicted world.

The results for the MAP inference experiments are shown in Figure 4 and the sizes of the respective models are in Table 2. In general, the possibilistic logic theories outperform MRFs for small evidence sets and do slightly worse for larger ones. Intuitively, the possibilistic logic theories approximate the density in a coarser way than the MRFs. This seems to lead to more robust results for hard problem instances (i.e. small evidence sizes), but less precise predictions for the easier instances. Interestingly, approximately compacting the possibilistic logic theories hardly affects the quality of the results in most cases. For NLTCS, however, the 10% theory has been reduced to the point that it only suggests to set everything to false, which is why the result in this case coincides with the baseline.

To evaluate the performance on marginal queries, we randomly sample conjunctions and compute the marginals of the conjunctions on the test set. Figure 5 shows the queries' predicted and empirical test-set probabilities for both possibilistic logic and MRFs on the NLTCS dataset. In this case, MRFs always obtained higher marginal

**Table 1.** **Left:** A possibilistic logic theory modelling a subset of the Plants dataset containing four states: California (ca), Montana (mt), New Mexico (nm) and Texas (tx). **Right:** An approximation of the possibilistic logic theory by short default rules (with antecedents being conjunctions of at most two literals).

| **Possibilistic logic theory** | **Approximation by default rules** |
|---|---|
| $(ca \vee nm \vee \neg tx \vee \neg mt, \lambda_{14}), (nm \vee \neg tx \vee \neg mt, \lambda_{13}), (ca \vee \neg tx \vee \neg mt, \lambda_{12}),$ | $\mid\!\sim \neg ca,\ \mid\!\sim \neg mt,\ \mid\!\sim \neg nm,\ \mid\!\sim \neg tx,\ nm \mid\!\sim tx,$ |
| $(ca \vee \neg nm \vee \neg mt, \lambda_{11}), (tx \vee \neg ca \vee \neg nm \vee mt, \lambda_{10}), (\neg ca \vee \neg tx \vee nm, \lambda_9),$ | $ca \wedge mt \mid\!\sim nm,\ ca \wedge mt \mid\!\sim tx,\ ca \wedge nm \mid\!\sim mt,$ |
| $(\neg ca \vee \neg nm \vee mt, \lambda_8), (\neg nm \vee tx \vee \neg mt, \lambda_7), (\neg ca \vee nm \vee \neg mt, \lambda_6),$ | $ca \wedge tx \mid\!\sim mt,\ ca \wedge tx \mid\!\sim nm,\ nm \wedge mt \mid\!\sim ca,$ |
| $(tx \vee \neg nm, \lambda_5), (\neg tx \vee \neg mt, \lambda_4), (\neg mt, \lambda_3), (\neg nm, \lambda_2), (\neg tx, \lambda_1), (\neg ca, \lambda_0)$ | $tx \wedge mt \mid\!\sim ca,\ tx \wedge mt \mid\!\sim nm$ |

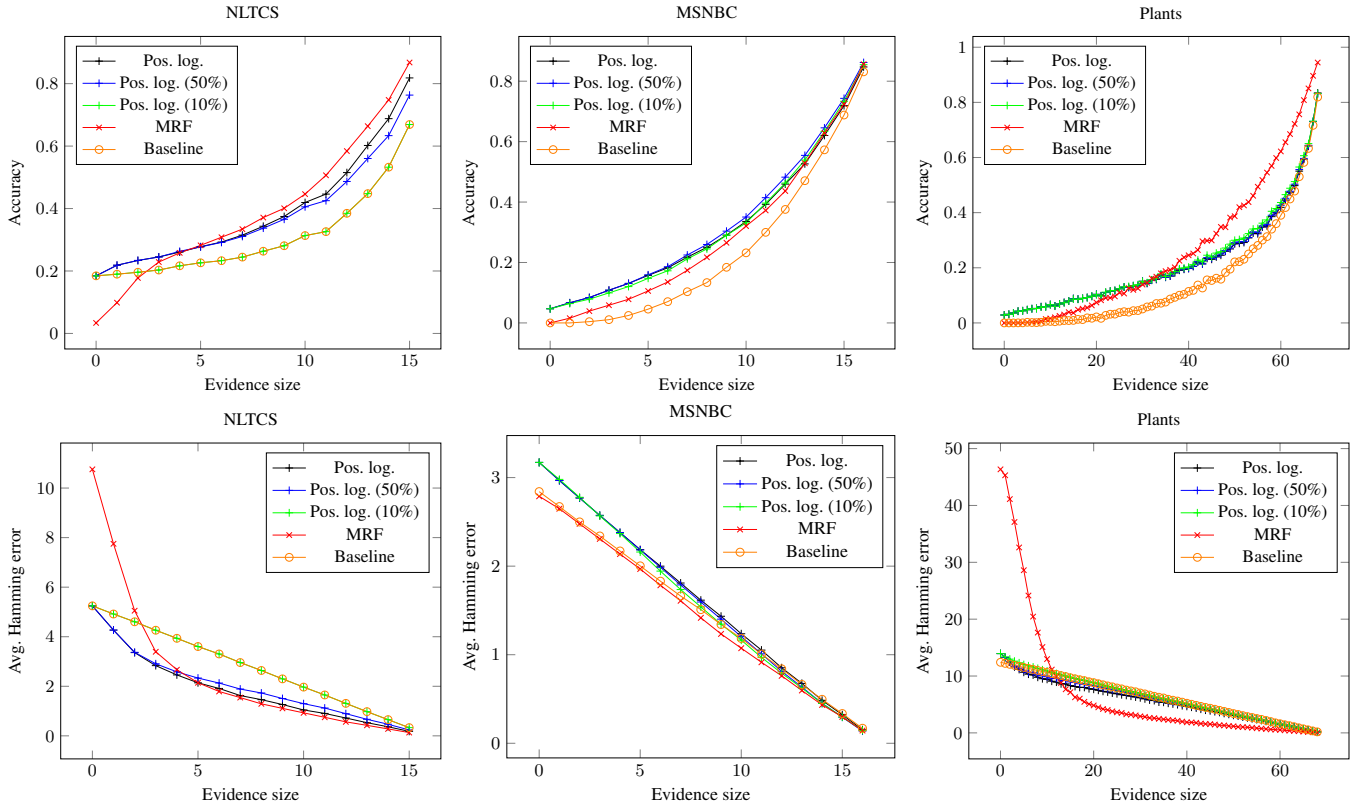



**Figure 4.** **Top:** Fraction of worlds correctly predicted by MAP inference on NLTCS, MSNBC and Plants datasets, measured on hold-out test sets. **Bottom:** Average Hamming error of possible worlds predicted by MAP inference, measured on hold-out test sets.

**Table 2.** The number of rules, average number of literals in a rule or branch for a tree, and the size which is the sum of rule lengths or number of nodes of a tree.

| **Dataset** | **Model** | **#Rules (#Branches)** | **Avg. lengh** | **Size** |
|---|---|---|---|---|
| **NLTCS** | PosLog | 121 | 3.0 | 363 |
| | PosLog (50%) | 71 | 2.4 | 172 |
| | PosLog (10%) | 16 | 1 | 16 |
| | Tree | 122 | 10.9 | 243 |
| | MRF | 135 | 1.9 | 254 |
| **MSNBC** | PosLog | 258 | 4.5 | 1153 |
| | PosLog (50%) | 172 | 3.3 | 572 |
| | PosLog (10%) | 53 | 2.1 | 109 |
| | Tree | 259 | 10.8 | 517 |
| | MRF | 136 | 1.9 | 289 |
| **Plants** | PosLog | 632 | 7.15 | 4523 |
| | PosLog (50%) | 467 | 4.8 | 2244 |
| | PosLog (10%) | 198 | 2.3 | 446 |
| | Tree | 655 | 18.6 | 1306 |
| | MRF | 2322 | 2.0 | 4713 |

log-likelihood. Nevertheless, the possibilistic logic theories' still offer competitive estimates along with the improved interpretability. The other datasets offer qualitatively similar results.

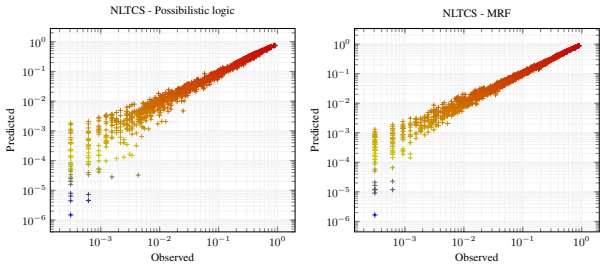



**Figure 5.** Scatter plots of estimated and empirical marginal probabilities of randomly generated queries on NLTCS dataset. **Left:** Possibilistic logic. **Right:** MRF.

# 6 CONCLUSIONS

We have introduced a practical method for constructing interpretable possibilistic logic models of probability distributions. The learned models support a variety of inference tasks, such as computing MAP queries and estimating marginal probabilities. Owing to the properties of possibilistic logic, the learned models can be easily edited by explicitly modifying, adding or removing logical rules, or they can be combined together. To maintain the ability to compute marginal probabilities after such modifications, we have proposed a parameter reestimation method based on geometric programming. Our experiments suggest that the method can be very useful for constructing interpretable logical theories from data.


# ACKNOWLEDGEMENTS

This work has been supported by a grant from the Leverhulme Trust (RPG-2014-164). Jesse Davis is partially supported by the KU Leuven Research Fund (C22/15/015), and FWO-Vlaanderen (G.0356.12, SBO-150033).


# A APPENDIX

## A.1 Probability-possibility transformation

A standard method to give a probabilistic interpretation to possibility degrees is by associating a possibility measure $\Pi$ with a family of probability measures, defined as $\mathcal{P}(\Pi) = \{P \,|\, P(A) \leq \Pi(A), \forall A \subseteq \Omega\}$. This view leads to the following probability-possibility transformation [11]. Let $p$ be a probability distribution on $\Omega = \{\omega_1, ..., \omega_n\}$ and assume w.l.o.g. that $p(\omega_i) \geq p(\omega_{i+1})$. Then $p$ induces a possibility distribution $\pi_p$ defined as $\pi_p(\omega_1) = 1$ and for $i > 1$:

$$\pi_p(\omega_i) = \begin{cases} \sum_{j=i}^{n} p(\omega_j) & \text{if } p(\omega_{i-1}) > p(\omega_i) \\ \pi_p(\omega_{i-1}) & \text{otherwise} \end{cases}$$

In [19], a syntactic counterpart of this transformation was used to associate each Markov logic network $M$ with a possibilistic logic theory $\Theta$, such that for $p$ the probability distribution associated with $M$ and $\pi$ the possibility distribution associated with $\Theta$ it holds that $p(\omega_1) \leq p(\omega_2)$ iff $\pi(\omega_1) \leq \pi(\omega_2)$. As a result, $(\Theta, \alpha) \vdash_{poss} \beta$ iff $\beta$ is true in all the most probable models of $\alpha$ (w.r.t. $p$). Other probabilistic interpretations of possibility distributions view possibility degrees as the contour function of a mass assignment, in the context of Dempster-Shafer evidence theory [28], or interpret possibility distributions as likelihood functions [9].

For completeness, we present a variant of Transformation 1 corresponding to a direct syntactic counterpart of the standard probability-possibility transformation.

**Transformation 4.** *Let $\mathcal{V}$ be the set of propositional variables. Let $T$ be a density estimation tree. Let $\mathcal{B} = \{B_1, \ldots, B_k\}$ be the set of all branches of the tree, represented as conjunctions, and let $p_1, \ldots, p_k$ be the estimated probabilities of worlds consistent with the respective branches, i.e. if $\omega \models B_i$ then $p(\omega) = p_i$. Let us assume w.l.o.g. that $p_i \geq p_{i+1}$. Let us define $w_1 = 1$. For $i > 1$, we define:*

$$w_i = \begin{cases} \sum_{j=i}^{n} p_j \cdot 2^{|\mathcal{V}|-|B_j|} & \text{if } p_{i-1} > p_i \\ w_{i-1} & \text{otherwise} \end{cases}$$

*We define the possibilistic logic theory corresponding to $T$ as $\Theta_T = \{(\neg B_i, 1 - w_i) | B_i \in \mathcal{B}\}$.*

## A.2 Lexicographic pruning of default rule theories

Default rules of the form $\alpha \mathrel{|\!\sim} \beta$, intuitively meaning "if $\alpha$ then typically $\beta$", offer a convenient way to make what is encoded by a possibilistic logic theory more explicit. If we take a purely qualitative view of possibilistic logic theories (i.e. if we see the weights merely as a way of specifying a ranking of possible worlds), a possibilistic logic theory $\Theta$ can be seen as a compact encoding of a set of default rules, i.e. a default theory, where a default $\alpha \mathrel{|\!\sim} \beta$ is in that theory if and only if $(\Theta, \{\alpha\}) \vdash_{poss} \beta$. The resulting default rules tend to be easy to interpret, but an exhaustive enumeration of all defaults would lead to theories in which many of the defaults are redundant. To cope with this problem, we rely on the *lexicographic closure* of default rules, which we describe next. The lexicographic closure [1] is one of several closures that have been studied in the field of non-monotonic reasoning, which allow us to represent an exhaustive set of defaults by a smaller set of defaults from which the complete set can be reconstructed. Smaller sets are usually easier for humans to understand.

To describe the lexicographic closure of default rules, first recall the *Z-ordering* from [24]. A default $\alpha \mathrel{|\!\sim} \beta$ is said to be *tolerated* by a set of defaults $\gamma_1 \mathrel{|\!\sim} \delta_1, ..., \gamma_m \mathrel{|\!\sim} \delta_m$ if the classical formula $\alpha \wedge \beta \wedge \bigwedge_i (\neg\gamma_i \vee \delta_i)$ is consistent. The *Z-ordering* is a stratification $\Delta_1, ..., \Delta_k$ of a set $\Delta$ of default rules, where each $\Delta_j$ contains all defaults $\alpha \mathrel{|\!\sim} \beta$ from $\Delta \setminus (\Delta_1 \cup ... \Delta_{j-1})$ which are tolerated by $\Delta \setminus (\Delta_1 \cup ... \cup \Delta_{j-1})$. It can be shown that such a stratification always exists when $\Delta$ satisfies some natural consistency properties (see [24] for details). Intuitively, $\Delta_1$ contains the most general default rules, $\Delta_2$ contains exceptions to the rules in $\Delta_1$, $\Delta_3$ contains exceptions to the rules in $\Delta_2$, etc. The lexicographic closure of a set of default rules is given as follows [1, 2]. For a possible world $\omega$, we write $sat(\omega, \Delta_j)$ for the number of defaults from $\Delta_j$ that are satisfied by $\omega$, i.e. $sat(\omega, \Delta_j) = |\{\alpha \mathrel{|\!\sim} \beta \colon (\alpha \mathrel{|\!\sim} \beta) \in \Delta_j, \omega \models \neg\alpha \vee \beta\}|$. We say that an interpretation $\omega_1$ is lex-preferred over an interpretation $\omega_2$, written $\omega_1 \prec \omega_2$, if there exists a $j$ such that $sat(\omega_1, \Delta_j) > sat(\omega_2, \Delta_j)$ while $sat(\omega_1, \Delta_i) = sat(\omega_2, \Delta_i)$ for all $i > j$. The default $\alpha \mathrel{|\!\sim} \beta$ is in the lex.closure of $\Delta$ if $\beta$ is satisfied in all the most lex-preferred models of $\alpha$, i.e. $\forall \omega \in [\![\alpha]\!] : (\omega \not\models \beta) \Rightarrow \exists \omega' \in [\![\alpha]\!] : \omega' \prec \omega$, where $[\![\alpha]\!]$ is the set of models of $\alpha$.

Now we can describe pruning of a (large) set of default rules $\Delta$, closed under the axioms of System P and rational monotonicity [18]. Our aim is not to construct the smallest set of defaults, as such a set could actually be more difficult to interpret. In particular, to maintain interpretability we only remove a rule if it is "implied" by a set of rules which are all shorter or equally long (as shorter rules are more interpretable). Furthermore note that methods for constructing the smallest set of defaults are likely to be computationally harder. Let $\Delta_1, \Delta_2, \ldots, \Delta_k$ be the Z-ordering of $\Delta$. Let us write $|\alpha \mathrel{|\!\sim} \beta|$ for the length of the default $\alpha \mathrel{|\!\sim} \beta$, e.g. the sum of literal occurrences in the antecedent $\alpha$ and consequent $\beta$. First we iteratively prune rules which are redundant in the rational closure sense – we remove a rule $\alpha \mathrel{|\!\sim} \beta$ if $(\Theta_R, \{\alpha\}) \vdash_{poss} \beta$ where $\Theta_R = \bigcup_{i=1}^{k} \{(\neg\gamma \vee \delta, 1/(k-i+1)) | \gamma \mathrel{|\!\sim} \delta \in \Delta_i\} \setminus \{\neg\alpha \vee \beta\}$. Then we iteratively prune the rules in $\Delta$ as follows. Iterating $i$ from 1 upwards, let $\alpha \mathrel{|\!\sim} \beta \in \Delta_i$. Let $L = |\alpha \mathrel{|\!\sim} \beta|$. Let $\Phi_j = \{\neg\gamma \vee \delta | \gamma \mathrel{|\!\sim} \delta \in \Delta_j \setminus \{\alpha \mathrel{|\!\sim} \beta\}$ and $|\gamma \mathrel{|\!\sim} \delta| \leq L$ and $\alpha \wedge (\neg\gamma \vee \delta) \not\vdash \bot\}$ and let $\Theta = \bigcup_{j=1}^{i-1} \{(\varphi, 1/(k-j+1)) | \varphi \in \Phi_j\}$ be a possibilistic logic theory. If $(\Theta, \{\alpha\}) \vdash_{poss} \beta$ we remove $\alpha \mathrel{|\!\sim} \beta$ from $\Delta$ and repeat this process for other rules in $\Delta$. It can be shown that all default rules from the initial set are contained in the lexicographic closure of the resulting, pruned set (although the closures themselves might differ if the original set was not closed).

## REFERENCES


[1] S. Benferhat, C. Cayrol, D. Dubois, J. Lang, and H. Prade, 'Inconsistency management and prioritized syntax-based entailment', in *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 93, pp. 640–645, (1993).

[2] Salem Benferhat, Jean F. Bonnefon, and Rui da Silva Neves, 'An overview of possibilistic handling of default reasoning, with experimental studies', *Synthese*, **146**(1-2), 53–70, (2005).

[3] Salem Benferhat, Didier Dubois, and Henri Prade, *Aggregation and Fusion of Imperfect Information*, chapter From Semantic to Syntactic Approaches to Information Combination in Possibilistic Logic, 141–161, Physica-Verlag HD, Heidelberg, 1998.

[4] D. Le Berre and A. Parrain, 'The SAT4J library, release 2.2.', *Journal on Satisfiability, Boolean Modeling and Computation*, **7**, 50–64, (2010).

[5] Stephen Boyd, Seung-Jean Kim, Lieven Vandenberghe, and Arash Hassibi, 'A tutorial on geometric programming', *Optimization and Engineering*, **8**(1), 67–127, (2007).

[6] Dennis L Bricker, Kenneth O Kortanek, and Lina Xu, 'Maximum likelihood estimates with order restrictions on probabilities and odds ratios: a geometric programming approach', *Advances in Decision Sciences*, **1**(1), 53–65, (1997).

[7] S. Della Pietra, V. Della Pietra, and J. Lafferty, 'Inducing features of random fields', **19**, 380–392, (1997).

[8] D. Dubois, J. Lang, and H. Prade, 'Possibilistic logic', in *Handbook of Logic in Artificial Intelligence and Logic Programming*, ed., D. Nute D. Gabbay, C. Hogger J. Robinson, volume 3, 439–513, Oxford University Press, (1994).

[9] D. Dubois, Serafn Moral, and Henri Prade, 'A semantics for possibility theory based on likelihoods', *Journal of Mathematical Analysis and Applications*, **205**(2), 359 – 380, (1997).

[10] D. Dubois and H. Prade, 'Possibility theory: qualitative and quantitative aspects', in *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, eds., D.M. Gabbay and P. Smets, volume 1, 169–226, Kluwer Academic, (1998).

[11] D. Dubois and Henri Prade, 'On several representations of an uncertain body of evidence', in *Fuzzy Information and Decision Processes*, eds., M.M. Gupta and E. Sanchez, 167–181, North-Holland, (1982).

[12] Didier Dubois, Laurent Foulloy, Gilles Mauris, and Henri Prade, 'Probability-possibility transformations, triangular fuzzy sets, and probabilistic inequalities', *Reliable Computing*, **10**(4), 273–297, (2004).

[13] F. Dupin de Saint-Cyr, J. Lang, and T. Schiex, 'Penalty logic and its link with Dempster-Shafer theory', in *Uncertainty in Artificial Intelligence*, pp. 204–211, (1994).

[14] Michael Grant and Stephen Boyd, 'Graph implementations for nonsmooth convex programs', in *Recent Advances in Learning and Control*, eds., V. Blondel, S. Boyd, and H. Kimura, Lecture Notes in Control and Information Sciences, 95–110, Springer-Verlag Limited, (2008).

[15] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx, March 2014.

[16] Wilhelmiina Hämäläinen and Matti Nykänen, 'Efficient discovery of statistically significant association rules', in *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pp. 203–212, (2008).

[17] Roberto J. Bayardo Jr. and Joseph Daniel Pehoushek, 'Counting models using connected components', in *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence, July 30 - August 3, 2000, Austin, Texas, USA.*, pp. 157–162, (2000).

[18] S. Kraus, D. Lehmann, and M. Magidor, 'Nonmonotonic reasoning, preferential models and cumulative logics', *Artif. Intelligence*, **44**(1-2), 167–207, (1990).

[19] O. Kuželka, J. Davis, and S. Schockaert, 'Encoding markov logic networks in possibilistic logic', in *Uncertainty in Artificial Intelligence, UAI*, (2015).

[20] Jérôme Lang, D. Dubois, and Henri Prade, 'A logic of graded possibility and certainty coping with partial inconsistency', in *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 188–196, (1991).

[21] D. Lowd and J. Davis, 'Improving markov network structure learning using decision trees', *Journal of Machine Learning Research*, **15**, 501–532, (2014).

[22] Daniel Lowd and Amirmohammad Rooshenas, 'The libra toolkit for probabilistic models', *CoRR*, **abs/1504.00110**, (2015).

[23] M. Meila and M. Jordan, 'Learning with mixtures of trees', **1**, 1–48, (2000).

[24] J. Pearl, 'System Z: A natural ordering of defaults with tractable applications to nonmonotonic reasoning', in *3rd Conference on Theoretical Aspects of Reasoning about Knowledge*, pp. 121–135, (1990).

[25] Parikshit Ram and Alexander G Gray, 'Density estimation trees', in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 627–635. ACM, (2011).

[26] Matthew Richardson and Pedro Domingos, 'Markov logic networks', *Machine Learning*, **62**(1-2), 107–136, (2006).

[27] Amirmohammad Rooshenas and Daniel Lowd, 'Learning sum-product networks with direct and indirect variable interactions', in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 710–718, (2014).

[28] G. Shafer, *A mathematical theory of evidence*, Princeton University Press, 1976.

[29] I-Cheng Yeh and Che-hui Lien, 'The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients', *Expert Syst. Appl.*, **36**(2), 2473–2480, (2009).

[30] L.A. Zadeh, 'Fuzzy sets as a basis for a theory of possibility', *Fuzzy Sets and Systems*, **1**, 3–28, (1978).